

SCinet Sandbox Project

**DICE/Obsidian
SCinet Sandbox Final Report**

10 May 2011

Prepared by



DICE Program Management Team



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

Table of Contents

1. Introduction	1
2. Project Description	1
2.1 Scope	1
2.2 Goals	1
2.3 Architecture	2
2.4 Participating Project Partners	3
3. Evaluation Activities	4
4. Evaluation Results	4
4.1 Local Testing at SC10	4
4.1.1 Local GPFS Testing	4
4.2 Testing From NASA GSFC	5
4.2.1 Single 10 GbE GPFS Testing	5
4.2.2 Four 10 GbE TCP/IP Testing Using Fusion-io Servers	7
5. Post SC10 Testing	9
6. Conclusions	11
6.1 Single 10 GbE GPFS Testing	11
6.2 Four 10 GbE TCP/IP Testing Using Fusion-io Servers	12
6.3 Post SC10 Testing - Effect of RTT	12
7. About Avetec and the DICE Program	12



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

1. Introduction

As part of a new direction for Supercomputing (SC) 2010, SCinet, the technical volunteer group that plans and supports the network infrastructure for the SC conferences, created a new initiative called the SCinet Sandbox (System Area Network Demonstration). This initiative was a combination of the previous Bandwidth Challenge and the Xnet program that SCinet conducted each year. The motivation behind the Sandbox program was to support strong technical evaluations and research demonstrations at the conference and also to involve SCinet technical professionals in the project and production of a final report documenting the project and all test results.

In August 2010, the DICE program along with its partner Obsidian Strategics submitted a proposal to the SCinet Sandbox team and its project was one of four accepted. The DICE SCinet Sandbox project was developed to understand and evaluate different data transfer and file system technologies in conjunction with wide area extended InfiniBand networks.

2. Project Description

Distributed computing and data management are two growing elements that are impacting the utilization of High Performance Computing (HPC). Both elements require high speed large bandwidth networks for large data transfers but also minimal latency to support small packet passing for operations like wide area file systems. These transfers must be secure and reliable, so encryption and reliable transit with minimal overhead is also needed in addition to performance. The SCinet Sandbox project sought to address this need.

2.1 Scope

This project focused on performance of applications over a wide area InfiniBand network from three separate sites to the conference floor at Supercomputing 2010 (SC10) in New Orleans. Technologies initially planned included Obsidian ES InfiniBand extenders, QDR/DDR/SDR InfiniBand, 10 and 100 GbE Ethernet LAN (local area network) and WAN (wide area network) technologies, wide area file system clients, pNFS technologies, InfiniBand attached storage, solid state disk technologies, various HPC I/O intensive applications, and several data transfer applications. The team developed tests to maximize the usage of all circuits utilizing various methods in the amount of time allotted for the testing by the SCinet Sandbox program.

2.2 Goals

The primary goal for this project was to do a full performance analysis of the Lustre and GPFS file systems over a distributed, nationwide, encrypted InfiniBand network. Standard I/O public domain performance evaluation tools, such as XDD, NUTTCP, NUTTSCP, dsync and IOzone, were used to establish baseline performance data. Where possible, a comparison was done using the InfiniBand infrastructure versus a standard 10 GbE implementation. The key InfiniBand product used throughout the testing was an Obsidian Longbow E100.

As a secondary goal, this project allowed vendors to participate with new technologies that are emerging or are in the final stages of development for release. This included Brocade's 100 GbE technology, DataDirect Network's SFA10K with QDR InfiniBand and SGI's Infinite Storage with DDR InfiniBand. In addition, this project allowed the research team and vendors to investigate the interoperability of these new technologies with a high speed wide area infrastructure.



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

2.3 Architecture

The DICE SCinet Sandbox project utilized resources on the SC10 showroom floor and from three remote sites: Avetec, located in Springfield, Ohio; NASA Goddard Space Flight Center (NASA GSFC), located in Greenbelt, Maryland; and Lawrence Livermore National Laboratory (LLNL), located in Livermore, California.

Each site was connected to the SC10 location by 10GbE connections, except for NASA GSFC which had the use of up to four 10 GbE connections and used Obsidian ES InfiniBand extenders to encapsulate IB traffic over the links.

High speed network connections were essential to the success of this project. The architecture included five Layer 2 wide area circuits and one routed circuit on the showroom floor from the SCinet infrastructure. ESnet provided one 10 GbE circuits and the remaining four were provided from National LambdaRail (NLR). The four NLR circuits completed connections to four circuits from NASA GSFC and a circuit to Avetec's USA40net, which shared one of the NLR paths with a NASA GSFC connection. The connection to NLR from these circuits was competed at the Starlight Exchange point in Chicago, Illinois.

A diagram of the architecture is shown in Figure 1.

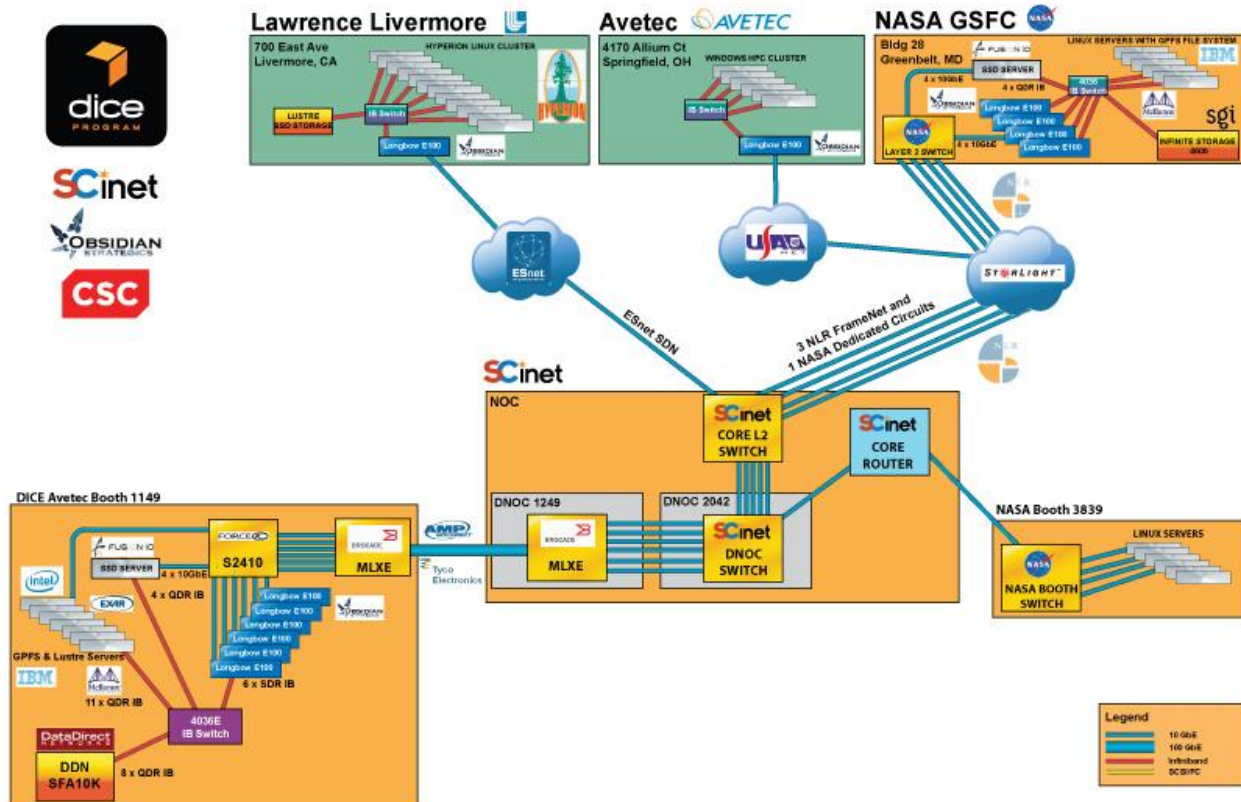


Figure 1: Architecture Diagram

(Locations with a green background were setup, but due to issues with equipment at those sites, no testing was completed.)



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

The three sites and several booths on the SC10 showroom floor contained various sets of hardware resources to accomplish the testing. The following tables describe the equipment that was planned and identifies the vendors involved.

Avetec Site

Purpose	Equipment	Provider
HPC Windows Cluster	Dell R610	Avetec
IB Extender	Obsidian Longbow E100	Obsidian
Disk Storage	RAID Xanadu 230	Avetec
IB switch	Mellanox IS 5030	Avetec

NASA GSFC Site

Purpose	Equipment	Provider
Servers	X8DTN+	Intel
IB attached Disk Storage	SGI Infinite Storage 4600	NASA - GSFC
Server	HP DL580	HP
SSD storage	ioDrive Octal	Fusion-io
IB switch	4036E	Mellanox
IB Extender	Obsidian Longbow E100	Obsidian

LLNL Site

Purpose	Equipment	Provider
Linux Cluster	Hyperion	LLNL
IB Extender	Obsidian Longbow E100	Obsidian
IB switch	Mellanox	LLNL

Avetec Booth SC10 Site

Purpose	Equipment	Provider
Servers	SuperMicro	Intel
IB attached Disk Storage	DDN SFA10000	DDN
Server	HP DL580	HP
SSD storage	ioDrive Octal	Fusion-io
IB switch	4036E	Mellanox
100 GbE Router	Brocade MLX	Brocade
10 GbE Switch	S2410	Force10
10 GbE PCIe Adapters	X3110 Single Port	Exar
QDR IB PCIe Adapters	ConnectX IB Dual Port	Mellanox
IB Extender	Obsidian Longbow E100	Obsidian

NASA Booth SC10 Site

Purpose	Equipment	Provider
Servers	SuperMicro	SuperMicro

SCinet DNOC at SC10 Site

Purpose	Equipment	Provider
100 GbE Router	Brocade MLX	Brocade

2.4 Participating Project Partners

The following vendors provided equipment and/or services in support of the SCinet Sandbox project: AMP NetConnect, Aspera, Brocade, DataDirect Networks, Exar, Force10, Fusion-io, IBM, Intel/SuperMicro, Mellanox, Microsoft, Obsidian Strategics, SGI and Tyco Electronics.

The following organizations participated in the project: Computer Sciences Corporation, ESnet, Lawrence Livermore National Lab (LLNL), NASA Goddard Space Flight Center (NASA GFSC), National Lambda Rail (NLR) and The Ohio State University.



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

This project included the involvement of several SCinet technical personnel. The following SCinet teams participated in this project for architecture support, test preparation, test execution and monitoring: SCinet Measurement, SCinet WAN, SCinet Fiber, SCinet Routing and SCinetOpenFabrics.

3. Evaluation Activities

The DICE SCinet Sandbox project was comprised of several tests to stress the wide area InfiniBand connections, benchmark wide area file system, and compare standard TCP/IP performance to InfiniBand data transfers, including those using RDMA. The DICE/Obsidian team set several goals for evaluating performance locally at the showroom floor and from the remote sites.

Locally at the showroom floor, I/O performance was to be measured for the local GPFS file systems.

Testing from the NASA GSFC site included different tests to measure performance for a remote GPFS file system connection. Four 10 GbE links were configured from the showroom floor from NASA GSFC. This configuration was crucial to providing the DICE team the ability to conduct larger than 10 Gbps performance transfer tests for RDMA over IB and standard TCP/IP. The following tests were to be conducted:

- Single 10 GbE GPFS
- Four 10 GbE TCP/IP using Fusion-io Servers

4. Evaluation Results

The DICE SCinet Sandbox project team completed several of the tests in the proposed test plan but ran into issues with equipment and circuit performance that forced cancellation and rescheduling or reduced performance for some tests. This report focuses on what the project team was able to accomplish. For example, after the close of the conference, the project team continued to investigate the effect of RTT on the performance of the different applications (see section 5).

4.1 Local Testing at SC10

Local testing at the SC10 showroom floor was conducted to provide a strong baseline for the wide area evaluations. The following sections detail the test configurations, parameters and benchmarks used to measure the local file systems performance.

4.1.1 Local GPFS Testing

The GPFS cluster on the SC10 showroom floor for the Linux testing consisted of four SuperMicro servers (nodes) connected to SFA10000 InfiniBand storage provided by DataDirect Networks. The cluster was configured as four Network Storage Devices (NSDs) with each NSD assigned to three 1.5 TB volumes for a total of twelve volumes and an 18 TB raw capacity file system. Two GPFS clients (SuperMicro servers) were local to the cluster on the showroom floor in New Orleans, and two clients were remotely located at NASA GSFC. The GPFS NSDs and clients were attached to the IB fabric with Mellanox QDR HCAs and Voltaire DDR HCAs respectively, while the Data Direct Network SFA10000 had four dedicated QDR connections for the NSDs.

Tests were conducted using IOzone and XDD benchmarks from the NSD and local clients. The following table describes each test and the associated result.



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

Benchmark	Servers Involved	# of Threads	Read Performance	Write Performance
IOZone	Single NSD GPFS Server	16	~2.9 GB/sec	~3.3 GB/sec
IOZone	Single DDR Client Server	16	~1.5 GB/sec	~1.9 GB/sec
XDD	1 st NSD GPFS Server	16	NA	~2.2 GB/sec
XDD	2 nd NSD GPFS Server	16	NA	~2.1 GB/sec
XDD	3 rd NSD GPFS Server	16	NA	~2.2 GB/sec
XDD	4 th NSD GPFS Server	16	NA	~1.3 GB/sec

Table 1: IOzone and XDD Benchmarks

The total aggregate write bandwidth with all four NSD servers on the XDD test was ~7.8 GB/sec. This was not quite 50% of the full line rate of the QDR links; however, the backend disk bandwidth was the reason for this limitation. Each NSD had 3 (8+1) volumes for a total of 24 disks. The average bandwidth per NSD of 1.95 GB/s divided by the 24 disks yields ~80 MB/s per disk that being SATA disks. This provided overall bandwidth was more than enough to support the wide area testing of a single SDR linked GPFS client.

4.2 Testing From NASA GSFC

Testing from NASA GSFC was conducted to provide additional wide area evaluations. The following sections detail the test configurations, parameters and benchmarks used to measure the remote file systems performance.

4.2.1 Single 10 GbE GPFS Testing

Part of the SCinet Sandbox testing at SC10 included demonstrating remote client access over the WAN to the earlier mentioned GPFS cluster on the SC10 showroom floor. Two GPFS clients (SuperMicro servers) clients were remotely located at NASA GSFC. On either end of the link from NASA GSFC to New Orleans was Obsidian Longbow hardware providing IB encapsulation. The link's theoretical peak from New Orleans to NASA GSFC was SDR InfiniBand based on the speed of the IB encapsulating hardware.

This test involved measurement of the wide area performance using the IOZone benchmark from a single DDR IB GPFS client at NASA GSFC using 16 threads. Tests were conducted over a two-hour span from the remote site and achieved up to 400-500 MB/sec read performance and 100-200 MB/sec write performance. The following graphs from the SCinet Measurement team servers depict the actual bandwidth utilization (Gbps) seen during that period.



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

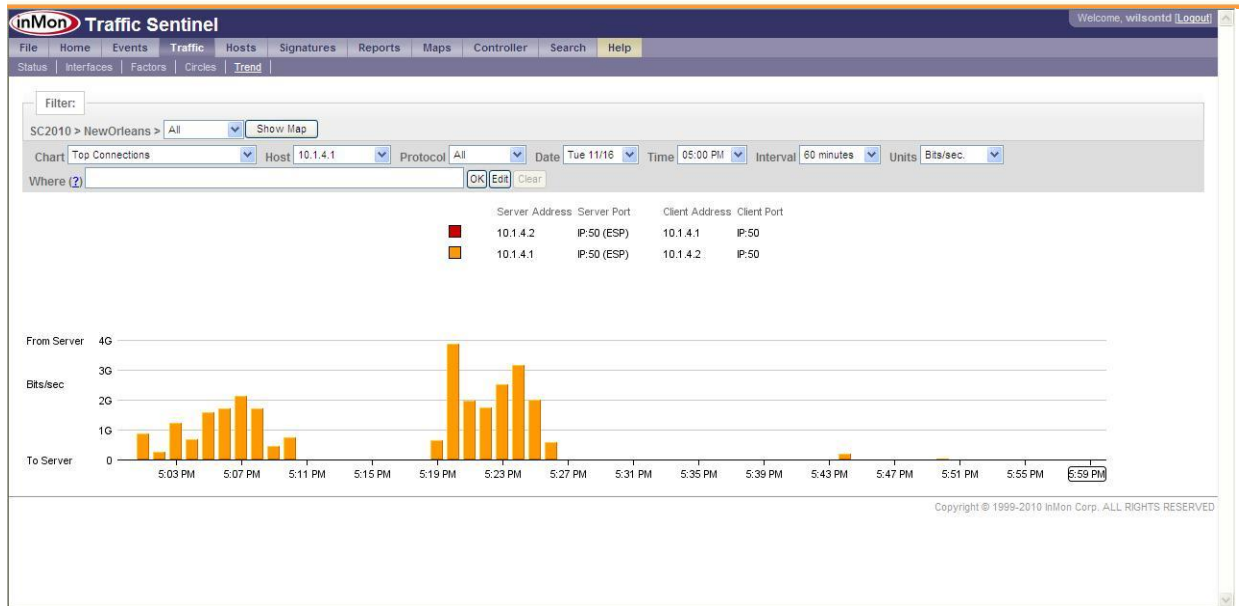


Figure 2: Combination of write and read testing with varying number of threads.

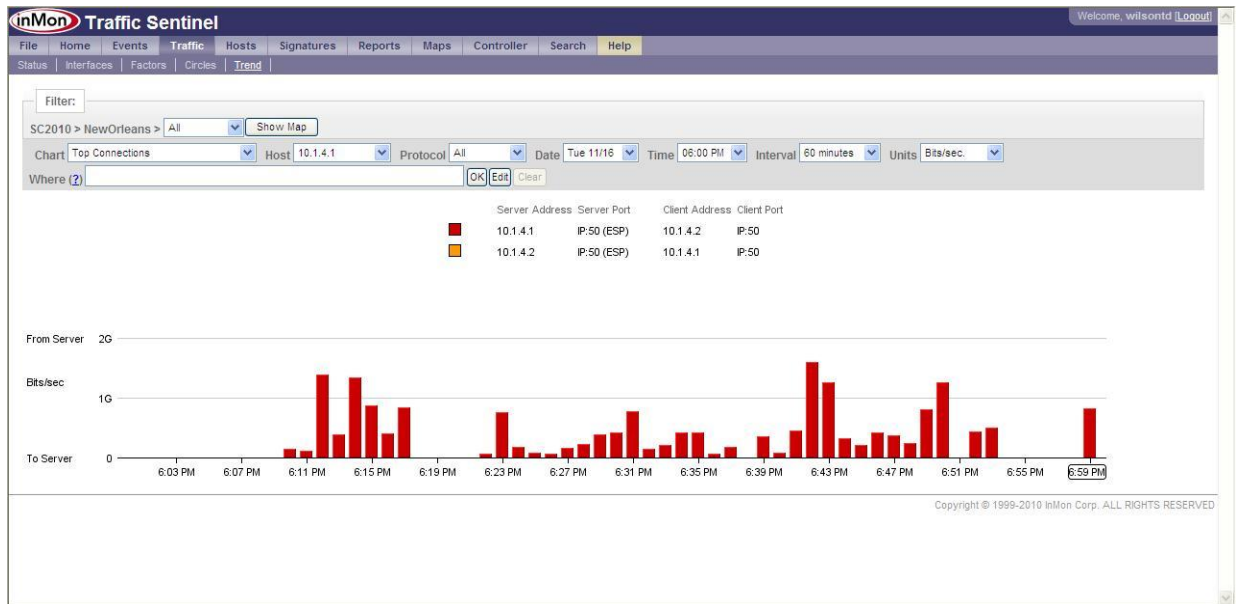


Figure 3: Write performance testing

Figure 2 is a combination of write and read testing with a varying number of threads. This variation results in the erratic nature in the range of 1 Gb/s to 4 Gb/s and everywhere in between. It was noted that the reads seemed to be pretty consistent when repeating a test, but writes, without changing parameters, seemed to have an erratic nature in themselves. In Figure 3 (later in the testing hour), the focus was more on trying to improve the write performance.

An important point to note was that the overall throughput on the Obsidian Longbows was purposely throttled back to half in the SDR link prior to testing in an attempt to circumvent the noise/congestion



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

performance degradation experienced on the wide area links. In theory ~500 MB/s was probably the best the team would see on either writes or reads. The graphs show that near theoretical wire speed could be achieved for reads; however, other unknown factors were limiting the write performance under 2 Gbit/s.

4.2.2 Four 10 GbE TCP/IP Testing Using Fusion-io Servers

Background and Configuration

In 2009 at SC09, NASA GSFC focused on single 10 GbE link performance testing. The emphasis shifted in 2010 to four links or 40 Mbps aggregate and the systems and applications that could move data at that kind of bandwidth. Besides the obvious challenge of securing four 10 GbE links, the hardware and software required to source and sync the data needed to be correspondingly robust. Leaning heavily on vendor loaners, NASA GSFC constructed the test infrastructure depicted in Figure 4. The configuration allowed testing of four native 10 GbE as well as encapsulated IB over IP as provided by the Obsidian Longbows, which also spread across the four links.

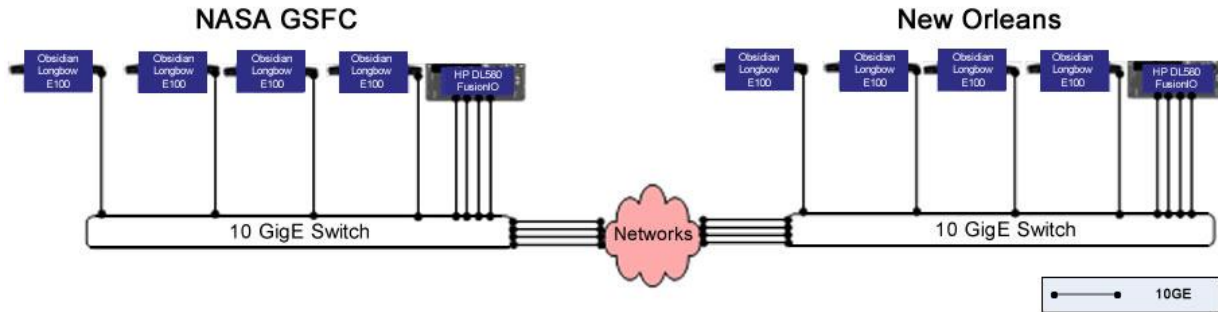


Figure 4: NASA GSFC Test Infrastructure

The infrastructure was anchored by two HP DL580 G7 servers, each configured as described in Table 2. One of the main selling points of the DL580 is the number and type of IO slots along with its ability to accommodate up to 4 CPUs and 64 memory DIMMs. The I/O configuration is not be understated given NASA GSFC’s objective of evaluating disk-to-disk data movement, not simply memory-to-memory. In rough terms, over 4000 MB/sec needed to be sustained throughout the data path to keep four links saturated. This implies high bandwidth storage at both ends. To meet the demands, Fusion-io provided two of their Octal cards, each capable of 6 GB/sec of bandwidth at a capacity of 5.12 TBs. The units supplied GSFC were configured with 1.2 TBs of storage. Two dual-ported 10GbE NICs and a dual-ported QDR InfiniBand card rounded out the hardware complement.

Item	Description
System/Motherboard	HP DL580G7
Processor	X7542 2.66 GHz (X2), six cores each
Memory	64 GB
Network Interface Card(s)	HP NC550SFP Dual Port 10 GbE Server Adapter (X2)
InfiniBand Host Card Adaptor	HP IB 4X QDR CX-2 PCI-e G2 Dual-port HCA
Solid State Disk Storage	Fusion-io Octal x16 PCI-E Gen2, 1.2 TB

Table 2: NASA GSFC Test Infrastructure



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

The software stack was built on Ubuntu 10.04 (lucid) as the selected Linux distribution. NASA GSFC sought compatible application that could stripe a single large file over the four links, either natively or via the Longbow connections. Very few packages were found to be multi-link aware in that fashion. The short list included the following packages:

- XDD, a rework of the popular storage benchmarking tool; upgrades being funded by Oak Ridge National Lab (ORNL) (<http://www.ioperformance.com/>)
- GridFTP, part of the Globus Toolkit (<http://globus.org/toolkit/>)
- Aspera, the lone commercial product from Aspera, Inc. (<http://www.asperasoft.com/>)
- dsync, developed by Obsidian Research (<http://www.obsidianresearch.com/>)

Test Results

Equipment was staged and configured at NASA GSFC prior to SC10. One of the HP DL580s was ultimately disconnected and sent to New Orleans. During the demonstration phase of the conference, link instability and limited test time (with all four links) curbed aspirations in terms of the objective goals NASA GSFC hoped to achieve. But subjectively, the time was quite successful. At various instances, key developers of the aforementioned applications were in the Avetec/DICE booth running tests, modifying configurations and writing code in real-time based on test results. A synopsis of the lessons learned is in the following table. They may not be surprising on an individual basis, but collectively they represent the number of items that have to be traded/balanced when constructing a high-bandwidth data transfer system:

Lesson	Learned
Link quality and RTT	The four links had different RTT values; three were in the mid-50s and one was in the 70s. Quality wise, one link was less stable than the other three. Overall, the Aspera product seemed more tolerant of the somewhat harsh network conditions.
IO type	Storage, in general, performs better with direct IO and multiple threads with asynchronous I/O being all the better. This is especially true with larger files. XDD and dsync offered more storage related tuning options which was key in terms of sourcing and syncing data and likewise keeping the network links fed. GridFTP and Aspera both have more storage IO improvements in the works.
File system	Hand in hand with IO type, this can be a facilitator or a bottleneck. Ability to pre-allocate file space is highly desirable. XDD performance, for instance, was dramatically better when xfs was used over ext2.
CPU core utilization	A feature found lacking was the ability to distribute application threads over all available cores and/or specific cores or sockets so as to take full advantage of both the processing power and memory bandwidth of the host server. For this to be truly effective, applications may need to spawn a user definable number of threads. Spreading and/or pinning interrupt processing to specific cores can also be advantageous. Some of this can be managed at the OS level already, but application awareness may be necessary as CPUs, memory and links all get faster.
Packet congestion	This one, at least in this case, was exclusive to the Obsidian dsync testing. Fanning and recombining data paths using one or two QDR IB ports versus four 10 GbE circuits led to notable inefficiencies even though the bandwidth overhead was clearly available.

Table 3: Lessons Learned



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

Representative, but not necessarily optimum (or maximum), performance data is illustrated in Figure 5. Captured using the inMon Traffic Sentinel, the graphs show a series of tests first exercising the individual links followed by two and four links. The application in this case was Aspera moving one, or in some cases, two files per link.

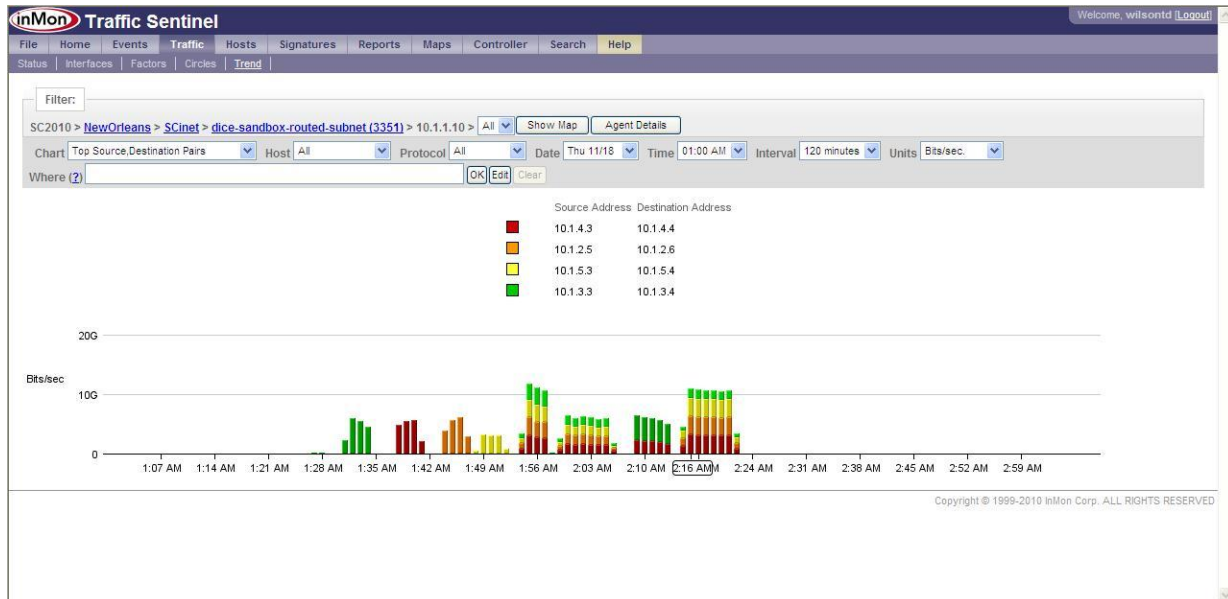


Figure 5: Example Application Performance Data

A significant block of time was spent working the Longbow dsync tests. In the end, the configuration was reduced to three links because one of the links proved to be too unstable. Even with the remaining three, the bandwidth per link had to be artificially limited to 6500 Mbps using an inherent feature of the Obsidian device. This helped mitigate the negative effects of the detected packet loss. Under these less than ideal conditions, dsync still managed to move a single large file in excess of 2000 MB/sec.

5. Post SC10 Testing

Not unexpectedly, the Sandbox project at SC10 left a number of questions unanswered. One of the primary questions unanswered was on the effect of RTT on the performance of the different applications. Another key question dealt with a nagging concern about the CPU-memory bandwidth of the HP DL580. Prior to returning the loaner equipment, additional testing was conducted at NASA GSFC.

The Longbow has the built-in capability to inject delay into the data path thereby simulating extended RTTs. At NASA GSFC, using the four available Longbows, two 10 GbE circuits were configured between the two DL580 servers. The following graph (Figure 6) shows the results for moving a single file of increasing sizes using dsync versus a varying RTT. It is unclear whether the fanning in effect at the smaller file size is real or if it is an anomaly in the way dsync calculates bandwidth. Regardless, the difference is relatively negligible.



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

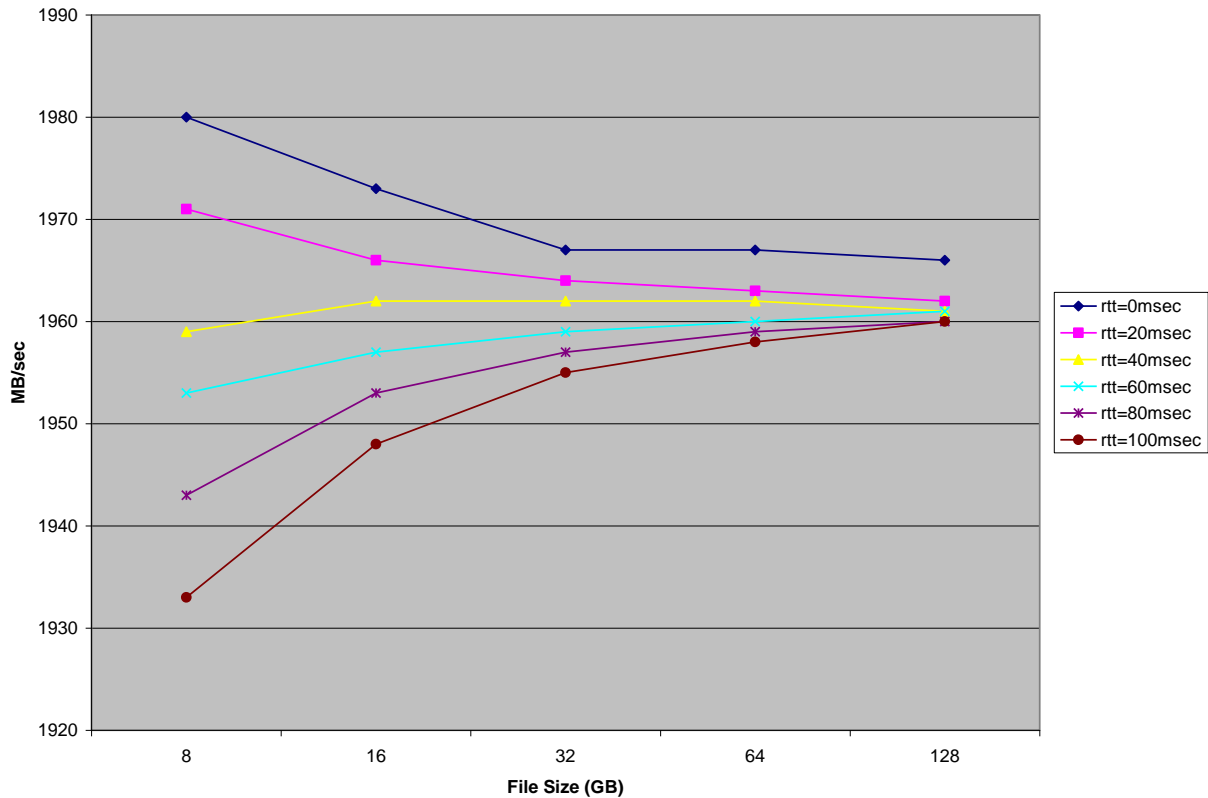


Figure 6: Obsidian Longbow Performance vs. RTT

NASA GSFC also conducted a more real-world test which involved moving multiple files and directories. Sixteen copies of a CentOS 5.5 installation disk (115104 files, 290 directories, 141GB) were aggregated on the source Fusion-io storage. At 0 msec RTT, the data was copied at 1430MB/sec, and at 100 msec RTT, it moved at 1381 MB/sec. One final dsync test was to the remove the Obsidian equipment all together and simply connect the DL580s back-to-back, QDR ports to QDR ports. In this configuration, dsync moved a single file at approximately 3100 MB/sec with 4000MB/sec being the theoretical max of the Gen 2 PCI Express x8 slot.

Early on, a concern developed that the overall performance of the four links, TCP/IP configuration was being limited to around 2500 MB/sec, independent of the number of active links. The DL580s were again hooked back-to-back but this time using the four 10 GbE connections. As a baseline, NUTTCP, a network performance tool, showed all four links capable of over 9800Mbps at 0 RTT. Then NASA GSFC ran the following series of tests using XDD, all at 0 RTT.

Links	8 GB	16 GB	32 GB	64 GB	128 GB
1	1231	1238	1239	1238	1237
2	2132	2114	2138	2140	2135
3	2403	2372	2383	2421	2437
4	2353	2353	2355	2398	2400

Table 4: Performance of the Four Links



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

Performance topped out at around 2400 MB/sec. The suspected culprit was CPU-memory bandwidth which effects the buffer copying required to feed the various interfaces. To substantiate the theory, multiple, different memory tests were run including stream, mbw and the home grown numa.c. All tests pointed to a peak CPU-memory bandwidth of around 2700 MB/sec, significantly less than the expected value for the installed X7542 CPUs. After verifying that Hemsphere mode was correctly enabled in the BIOS, the fact that the DL580 was effectively half-populated was assumed to be the limiting factor. The remedy was to combine the CPUs and memory of both DL580s into a single machine and rerun the memory tests. The measured bandwidth numbers remained unchanged. A subsequent meeting with HP uncovered an additional “green” related BIOS setting that could have been impacting performance. HP is in the process of setting up both a DL580 and DL980 for NASA GSFC to conduct additional testing for the purpose of exploring CPU-memory bandwidth as a function of socket count.

6. Conclusions

The project team had an aggressive plan to develop a nationwide test bed to evaluate various aspects of InfiniBand technology in geographically dispersed locations. The team worked closely with the SCinet team to place equipment in both the NASA booth and the Avetec/DICE booth and establish a 10 gigabit network between the booths while at the same time leveraging NLR and USA40net to interconnect to LLNL, NASA GSFC and Avetec. In addition to the aggressive hardware and network architecture, the team had an aggressive series of tests to evaluate the performance and capabilities.

This project allowed vendors and the project team to test and evaluate new technologies that are emerging or are in the final stages of development for release. The project team discovered that Brocade’s 100 GbE technology, DDN’s SFA10K with QDR InfiniBand and SGI’s Infinite Storage with DDR InfiniBand all worked as advertised and allowed for inter-operability. Clearly, the vendors had a commitment to meeting IB standards and were successful.

The project team’s utilization of the Obsidian Longbow product as the “anchor” technology proved to be a wise choice. The Longbow product met or exceeded the team’s expectations. The unit showed the same performance whether their encryption capabilities were turned on or off. For a full report on performance of the Longbow, the team encourages a review of the work of Dr. D.K. Panda at the Ohio State University (www.cse.ohio-state.edu/~panda or www.diceprogram.org).

Overall, the long haul network to the Avetec and NASA GSFC facilities were inconsistent at best and led the team to abbreviate the test plan and eliminate some of the tests. This report focuses on the tests that were able to be accomplished.

Throughout the testing, the vendors and the rest of the project team had opportunities to try various combinations of configurations features. This allowed everyone to learn and understand the impact of settings as they relate to particular applications.

6.1 Single 10 GbE GPFS Testing

Due to noise and congestion from all the traffic and activity at the SC conference, the project team needed to throttle back to half in the SDR link prior to testing. Because of the congestion, the team estimated that the best they could achieve was~500 MB/s on either of the writes or reads. The test proved that the numbers for reads would be consistently met. The issue with the performance on the writes could not be solved in the timetable of the conference. Significant bandwidth was being consumed by other tests running on the network. In the future, the hope is that SCinet will allow these major tests the option to schedule a high speed VPN. A consistent quality of service would allow the adjustment of



DATA INTENSIVE COMPUTING ENVIRONMENT

REAL TESTING | REAL DATA | REAL RESULTS

settings and configurations and enable a better understanding of the impact of these changes rather than wondering whether the changes were impacted by congestion. The local tests proved that the Obsidian Longbows were indeed meeting their performance expectations.

6.2 Four 10 GbE TCP/IP Testing Using Fusion-io Servers

Without much surprise, the testing revealed that storage tends to perform better with direct IO and multiple threads than with asynchronous IO. This was especially true with larger files. One of the disappointments was the inability to ensure broad CPU core utilization. There were no tools or methods to distribute application threads over all available cores and/or specific cores or sockets so as to take full advantage of both the processing power and memory bandwidth of the host server. This is hoped to be rectified soon. The project team did discover a sequence of events that lead to packet congestion. While doing extensive Obsidian dsync testing, the team discovered in some cases fanning and recombining data paths using one or two QDR IB ports versus four 10GbE circuits led to notable inefficiencies even though the bandwidth overhead was clearly available. The team was impressed with the Aspera product which was more tolerant of the harsh network conditions.

6.3 Post SC10 Testing- Effect of RTT

The post SC10 testing was challenging but created an opportunity to learn new information about BIOS configurations. One of the challenges in this test was that the CPU seemed to have an artificial limit on the amount of data that could be moved. In discussions with HP representatives, the team learned that there is a new setting in the BIOS that is directly related to a green initiative. The team is working with HP to deactivate this setting to see if that is reason for the apparent limit on data transfer rates.

7. About Avetec and the DICE Program

The DICE program is administered by Avetec, Inc., a non-profit public benefit research organization that uses virtual testing environments to help solve complex problems, such as those in aerospace engineering. Avetec's high-performance computing (HPC) Research Division – the Data Intensive Computing Environment – is a geographically dispersed test environment that conducts technology testing and validation for new and emerging HPC data management solutions. The DICE team works with the HPC industry, data centers (government and industry) and the research community to evaluate new and emerging products and technology that enhance research computing data and results throughput.

This project was funded in part by the Department of Energy (DOE) Sandia Labs.